

# QUADRATIC MIXED FINITE ELEMENT APPROXIMATIONS OF THE MONGE-AMPÈRE EQUATION IN 2D

GERARD AWANOU

**ABSTRACT.** We give error estimates for a mixed finite element approximation of the two-dimensional elliptic Monge-Ampère equation with the unknowns approximated by Lagrange finite elements of degree two. The variables in the formulation are the scalar variable and the Hessian matrix.

## 1. INTRODUCTION

Let  $\Omega$  be a convex polygonal domain of  $\mathbb{R}^2$  with boundary  $\partial\Omega$ . We are interested in a mixed finite element method for the nonlinear elliptic Monge-Ampère equation: find a smooth convex function  $u$  such that

$$(1.1) \quad \begin{aligned} \det(D^2u) &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega. \end{aligned}$$

For  $u \in C^2(\Omega)$ ,  $D^2u = \left( (\partial^2 u) / (\partial x_i \partial x_j) \right)_{i,j=1,\dots,2}$  denotes the Hessian matrix of  $u$  and  $\det D^2u$  denotes its determinant. The function  $f$  defined on  $\Omega$  is assumed to satisfy  $f \geq c_0 > 0$  for a constant  $c_0 > 0$  and we assume that  $g \in C(\partial\Omega)$  can be extended to a function  $\tilde{g} \in C(\overline{\Omega})$  which is convex in  $\Omega$ .

We consider a mixed formulation with unknowns the scalar variable  $u$  and the Hessian  $D^2u$ . The scalar variable and the components of the Hessian are approximated by Lagrange elements of degree  $k \geq 2$ . The method considered in this paper was analyzed from different point of views in [9] and [4] for smooth solutions of (1.1). In both [9] and [4] the convergence of the method for Lagrange elements of degree  $k = 1$  and  $k = 2$  was left unresolved. In this paper we resolve this issue for quadratic elements.

The ingredients of our approach consist in a fixed point argument, which yields the convergence of a time marching method, a "rescaling argument", i.e. the solution of a rescaled version of the equation, and the continuity of the eigenvalues of a matrix as a function of its entries. This is the same approach we took in the case of the standard finite element discretization of the Monge-Ampère equation [3].

With the mixed methods, one can apply directly Newton's method to the discrete nonlinear problem and still have numerical evidence of convergence to a larger class of non smooth solutions than what is possible with the standard finite element discretization. We refer to [9, 8] for the numerical results. Moreover with the standard finite element discretization [3], convexity must be enforced weakly through appropriate iterative methods. Although the number of unknowns in the mixed methods is

---

The author was partially supported by NSF DMS grant No 1319640.

higher, in [9, 8] the discrete Hessian was eliminated from the discrete equations in the implementation. However, as observed in [4] this prevents numerical convergence for smooth solutions when linear elements are used to approximate all the unknowns. We note that in [9] a stabilized method was proposed which works numerically for non smooth solutions in two dimension. It consists in using piecewise constants for the discrete Hessian and linear elements for the scalar variable. The analysis for smooth solutions of the lowest order methods discussed in [4, 9] cannot be done with the approach of this paper. The techniques used in this paper generalize to the three-dimensional problem but only for  $k \geq 3$ . It should be possible to extend the approach taken in this paper to the formulation where discontinuous elements are used to approximate the unknowns [9]. Numerical results reported in [9] indicate the latter approach could lead to a less accurate approximation of the Hessian. For simplicity, and to focus on the methodology we present, we do not consider such an extension in this paper.

We organize the paper as follows. In the second section we introduce some notation and preliminaries. The error analysis of the mixed method is done in section 3.

## 2. NOTATION AND PRELIMINARIES

We use the usual notation  $L^p(\Omega)$ ,  $2 \leq p \leq \infty$  for the Lebesgue spaces and  $H^s(\Omega)$ ,  $1 \leq s < \infty$  for the Sobolev spaces of elements of  $L^2(\Omega)$  with weak derivatives of order less than or equal to  $s$  in  $L^2(\Omega)$ . We recall that  $H_0^1(\Omega)$  is the subset of  $H^1(\Omega)$  of elements with vanishing trace on  $\partial\Omega$ . We also recall that  $W^{s,\infty}(\Omega)$  is the Sobolev space of functions with weak derivatives of order less than or equal to  $s$  in  $L^\infty(\Omega)$ . For a given normed space  $X$ , we denote by  $X^2$  the space of vector fields with components in  $X$  and by  $X^{2 \times 2}$  the space of matrix fields with each component in  $X$ .

The norm in  $X$  is denoted by  $\|\cdot\|_X$  and we omit the subscript  $\Omega$  and superscripts 2 and  $2 \times 2$  when it is clear from the context. The inner product in  $L^2(\Omega)$ ,  $L^2(\Omega)^2$ , and  $L^2(\Omega)^{2 \times 2}$  is denoted by  $(\cdot, \cdot)$  and we use  $\langle \cdot, \cdot \rangle$  for the inner product on  $L^2(\partial\Omega)$  and  $L^2(\partial\Omega)^2$ . For inner products on subsets of  $\Omega$ , we will simply append the subset notation.

We denote by  $n$  the unit outward normal vector to  $\partial\Omega$ . We recall that for a matrix  $A$ ,  $A_{ij}$  denote its entries and the cofactor matrix of  $A$ , denoted  $\text{cof } A$ , is the matrix with entries  $(\text{cof } A)_{ij} = (-1)^{i+j} \det(A)_i^j$  where  $\det(A)_i^j$  is the determinant of the matrix obtained from  $A$  by deleting its  $i$ th row and its  $j$ th column. For two matrices  $A = (A_{ij})$  and  $B = (B_{ij})$ ,  $A : B = \sum_{i,j=1}^2 A_{ij} B_{ij}$  denotes their Frobenius inner product. A quantity which is constant is simply denoted by  $C$ .

For a scalar function  $v$  we denote by  $Dv$  its gradient vector and recall that  $D^2v$  denotes the Hessian matrix of second order derivatives. The divergence of a matrix field is understood as the vector obtained by taking the divergence of each row.

In this section and section 3 we assume that (1.1) has a solution which is sufficiently smooth. Put  $\sigma = D^2u$ . Then the unique convex solution  $u \in H^3(\Omega)$  of (1.1) satisfies

the following mixed problem: Find  $(u, \sigma) \in H^2(\Omega) \times H^1(\Omega)^{2 \times 2}$  such that

$$(2.1) \quad \begin{aligned} (\sigma, \tau) + (\operatorname{div} \tau, Du) - \langle Du, \tau n \rangle &= 0, \forall \tau \in H^1(\Omega)^{2 \times 2} \\ (\det \sigma, v) &= (f, v), \forall v \in H_0^1(\Omega) \\ u &= g \text{ on } \partial\Omega. \end{aligned}$$

It is proved in [4] that the above variational problem is well defined.

**2.1. Discrete variational problem.** We denote by  $\mathcal{T}_h$  a triangulation of  $\Omega$  into simplices  $K$  and assume that  $\mathcal{T}_h$  is quasi-uniform. We denote by  $V_h$  the standard Lagrange finite element space of degree  $k \geq 2$  and denote by  $\Sigma_h$  the space of symmetric matrix fields with components in the Lagrange finite element space of degree  $k \geq 2$ . Let  $I_h$  denote the standard Lagrange interpolation operator from  $H^s(\Omega)$ ,  $s \geq k+1$  into the space  $V_h$ . We use as well the notation  $I_h$  for the matrix version of the Lagrange interpolation operator mapping  $H^s(\Omega)^{2 \times 2}$ , for  $s \geq k+1$ , into  $\Sigma_h$ . We consider the problem: find  $(u_h, \sigma_h) \in V_h \times \Sigma_h$  such that

$$(2.2) \quad \begin{aligned} (\sigma_h, \tau) + (\operatorname{div} \tau, Du_h) - \langle Du_h, \tau n \rangle &= 0, \forall \tau \in \Sigma_h \\ (\det \sigma_h, v) &= (f, v), \forall v \in V_h \cap H_0^1(\Omega) \\ u_h &= g_h \text{ on } \partial\Omega, \end{aligned}$$

where  $g_h = I_h \tilde{g}$ . It follows from the analysis in [9, 4] that (2.2) is well-posed for  $k \geq 3$  and error estimates were given. In section 3 we give an error analysis valid for  $k \geq 2$ . For  $v_h \in V_h$ , we will make the abuse of notation of using  $D^2 v_h$  to denote the Hessian of  $v_h$  computed element by element. We will need the broken Sobolev norm

$$\|v\|_{H^k(\mathcal{T}_h)} = \left( \sum_{K \in \mathcal{T}_h} \|v\|_{H^k(K)}^2 \right)^{\frac{1}{2}}.$$

**2.2. Properties of the Lagrange finite element spaces.** We recall some properties of the Lagrange finite element space of degree  $k \geq 1$  that will be used in this paper. They can be found in [7, 5]. We have

Interpolation error estimates.

$$(2.3) \quad \begin{aligned} \|v - I_h v\|_{H^j} &\leq Ch^{k+1-j} \|v\|_{H^{k+1}}, \forall v \in H^s(\Omega), j = 0, 1, \\ \|v - I_h v\|_{L^\infty} &\leq Ch^k \|v\|_{H^{k+1}}, \forall v \in H^s(\Omega). \end{aligned}$$

Inverse inequalities

$$(2.4) \quad \|v\|_{L^\infty} \leq Ch^{-1} \|v\|_{L^2}, \forall v \in V_h$$

$$(2.5) \quad \|v\|_{H^1} \leq Ch^{-1} \|v\|_{L^2}, \forall v \in V_h$$

$$(2.6) \quad \|v\|_{H^{k+1}(\mathcal{T}_h)} \leq Ch^{-k-1} \|v\|_{L^2}, \forall v \in V_h.$$

Scaled trace inequality

$$(2.7) \quad \|v\|_{L^2(\partial\Omega)} \leq Ch^{-\frac{1}{2}} \|v\|_{L^2}, \forall v \in V_h.$$

**2.3. Algebra with matrix fields.** We collect in the following lemma some properties of matrix fields, the proof of which can be found in [4, 1].

**Lemma 2.1.** *For  $K \in \mathcal{T}_h$  and  $u, v \in C^2(K)$  we have*

$$(2.8) \quad \det D^2 u - \det D^2 v = \operatorname{cof}(tD^2 u + (1-t)D^2 v) : (D^2 u - D^2 v),$$

for some  $t \in [0, 1]$ . It can be shown that  $t = 1/2$ , [6].

For two  $2 \times 2$  matrix fields  $\eta$  and  $\tau$

$$(2.9) \quad \|\operatorname{cof}(\eta) : \tau\|_{L^2} \leq C \|\eta\|_{L^\infty} \|\tau\|_{L^2},$$

$$(2.10) \quad \operatorname{cof}(\eta) - \operatorname{cof}(\tau) = \operatorname{cof}(\eta - \tau).$$

**2.4. Continuity of the eigenvalues of a matrix as a function of its entries.**

Let  $\lambda_1(A)$  and  $\lambda_2(A)$  denote the smallest and largest eigenvalues of the symmetric matrix  $A$ . We have

**Lemma 2.2** ([3], Lemma 3.1). *There exists constants  $m, M > 0$  independent of  $h$  and a constant  $C_{conv} > 0$  independent of  $h$  such that for all  $v_h \in V_h$  with  $v_h = g_h$  on  $\partial\Omega$  and*

$$\|v_h - I_h u\|_{H^1} < C_{conv} h^2,$$

we have

$$m \leq \lambda_1(\operatorname{cof} D^2 v_h(x)) \leq \lambda_2(\operatorname{cof} D^2 v_h(x)) \leq M, \forall x \in K, K \in \mathcal{T}_h.$$

The following lemma was used implicitly in [1, 3, 2].

**Lemma 2.3.** *Assume  $0 < \alpha < 1$  and  $\alpha \leq (m + M)/(2m)$  for constants  $m, M > 0$ . Let  $B$  be a symmetric matrix field such that*

$$0 < m\alpha \leq \lambda_1(B(x)) \leq \lambda_2(B(x)) \leq M\alpha, \forall x \in \Omega.$$

Then for  $\nu = (m + M)/2$

$$\gamma \equiv \sup_{\substack{v, w \in V_h \\ |v|_{H^1}=1, |w|_{H^1}=1}} \left| (Dv, Dw) - \frac{1}{\nu} (BDv, Dw) \right|,$$

satisfies  $0 < \gamma < 1$ .

*Proof.* Since  $\lambda_1(B)$  and  $\lambda_2(B)$  are the minimum and maximum respectively of the Rayleigh quotient  $((Bz) \cdot z)/\|z\|^2$ , where  $\|z\|$  denotes the Euclidean norm of  $\mathbb{R}^2$ , we have for  $x \in \Omega$

$$m\alpha \|z\|^2 \leq (B(x)z) \cdot z \leq M\alpha \|z\|^2, z \in \mathbb{R}^2.$$

This implies

$$m\alpha |w|_{H^1}^2 \leq \int_{\Omega} [B(x)Dw(x)] \cdot Dw(x) \, dx \leq M\alpha |w|_{H^1}^2, w \in V_h.$$

If we assume in addition that  $|w|_{H^1} = 1$ , we get

$$m\alpha \leq \int_{\Omega} [B(x)Dw(x)] \cdot Dw(x) \, dx \leq M\alpha, w \in V_h.$$

It follows that

$$(1 - \frac{M\alpha}{\nu}) \leq \int_{\Omega} [I - \frac{1}{\nu} B(x) Dw(x)] \cdot Dw(x) dx \leq (1 - \frac{m\alpha}{\nu}), w \in V_h.$$

Since  $\nu = (m + M)/2$ , we have

$$1 - \frac{\alpha M}{\nu} = \frac{m + M - 2M\alpha}{m + M} < 1$$

$$1 - \frac{\alpha m}{\nu} = \frac{m + M - 2m\alpha}{m + M} < 1.$$

If we define

$$\beta \equiv \sup_{v \in V_h, |v|_{H^1}=1} \left| (Dv, Dv) - \frac{1}{\nu} (BDv, Dv) \right|,$$

by the assumptions on  $\alpha$ , we have

$$0 < \beta < 1.$$

We can define a bilinear form on  $V_h$  by the formula

$$(p, q) = \int_{\Omega} [(I - \frac{1}{\nu} B(x)) Dp(x)] \cdot Dq(x) dx.$$

Then because

$$(p, q) = \frac{1}{4} ((p + q, p + q) - (p - q, p - q)),$$

and using the definition of  $\beta$ , we get assuming that  $|p|_{H^1} = |q|_{H^1} = 1$ ,

$$\begin{aligned} |(p, q)| &\leq \frac{\beta}{4} (p + q, p + q) + \frac{\beta}{4} (p - q, p - q) \\ &\leq \frac{\beta}{4} |p + q|_{H^1}^2 + \frac{\beta}{4} |p - q|_{H^1}^2 = \beta. \end{aligned}$$

This completes the proof. □

### 3. ERROR ANALYSIS OF THE MIXED METHOD FOR SMOOTH SOLUTIONS

We will assume without loss of generality that  $h \leq 1$ . The goal of this section is to prove the local solvability of (2.2) for Lagrange elements of degree  $k \geq 2$ . We define for  $\rho > 0$ ,

$$\bar{B}_h(\rho) = \{(w_h, \eta_h) \in V_h \times \Sigma_h, \|w_h - I_h u\|_{H^1} \leq \rho, \|\eta_h - I_h \sigma\|_{L^2} \leq h^{-1} \rho\}.$$

We are interested in elements  $(w_h, \eta_h) \in V_h \times \Sigma_h$  satisfying

$$(3.1) \quad (\eta_h, \tau) + (\operatorname{div} \tau, Dw_h) - \langle Dw_h, \tau n \rangle = 0, \forall \tau \in \Sigma_h.$$

We define

$$Z_h = \{(w_h, \eta_h) \in V_h \times \Sigma_h, w_h = g_h \text{ on } \partial\Omega, (w_h, \eta_h) \text{ solves (3.1)}\} \text{ and}$$

$$B_h(\rho) = \bar{B}_h(\rho) \cap Z_h.$$

In [4] the local solvability of (2.2) was obtained by a fixed point argument which consists in a linearization at the exact solution of (1.1). To be able to obtain results for quadratic elements we use a time marching method combined with a rescaling

argument. This is the point of view we took in [3, 2]. We first describe the time marching method at the continuous level.

Let  $\nu > 0$ . We consider the sequence of problems

$$\begin{aligned} -\nu \Delta u^{r+1} &= -\nu \Delta u^r + \det D^2 u^r - f \text{ in } \Omega \\ u^{r+1} &= g \text{ on } \partial\Omega. \end{aligned}$$

Put  $\sigma^{r+1} = D^2 u^{r+1}$ . We obtain the equivalent problems

$$\begin{aligned} \sigma^{r+1} &= D^2 u^{r+1} \text{ in } \Omega \\ -\nu \operatorname{tr} \sigma^{r+1} &= -\nu \operatorname{tr} \sigma^r + \det \sigma^r - f, \text{ in } \Omega \\ u^{r+1} &= g \text{ on } \partial\Omega, \end{aligned}$$

where  $\operatorname{tr} A$  denotes the trace of the matrix  $A$ .

We are thus lead to consider the sequence of discrete problems: find  $(u_h^{r+1}, \sigma_h^{r+1}) \in V_h \times \Sigma_h$  such that  $u_h^{r+1} = g_h$  on  $\partial\Omega$  and

$$(3.2) \quad (\sigma_h^{r+1}, \tau) + (\operatorname{div} \tau, D u_h^{r+1}) - \langle D u_h^{r+1}, \tau n \rangle = 0, \forall \tau \in \Sigma_h$$

$$(3.3) \quad -\nu (\operatorname{tr} \sigma^{r+1}, v) = -\nu (\operatorname{tr} \sigma^m, v) + (\det \sigma_h^r - f, v), \forall v \in V_h \cap H_0^1(\Omega),$$

given an initial guess  $(u_h^0, \sigma_h^0)$ . We prove below the convergence of  $(u_h^{r+1}, \sigma_h^{r+1})$  to a local solution  $(u_h, \sigma_h)$  of the discrete problem (2.2). Although (3.2)–(3.3) may be used in the computations, it is better to use in practice Newton's method.

Let  $\alpha > 0$ . We define a mapping  $T : V_h \times \Sigma_h \rightarrow V_h \times \Sigma_h$  by

$$T(w_h, \eta_h) = (T_1(w_h, \eta_h), T_2(w_h, \eta_h)),$$

where  $T_1(w_h, \eta_h)$  and  $T_2(w_h, \eta_h)$  satisfy

$$\begin{aligned} (\eta_h - T_2(w_h, \eta_h), \tau) + (\operatorname{div} \tau, D(w_h - T_1(w_h, \eta_h))) \\ - \langle D(w_h - T_1(w_h, \eta_h)), \tau n \rangle = (\eta_h, \tau) \\ + (\operatorname{div} \tau, D w_h) - \langle D w_h, \tau n \rangle, \quad \forall \tau \in \Sigma_h \end{aligned} \quad (3.4)$$

$$(3.5) \quad -\nu (\operatorname{tr} T_2(w_h, \eta_h), v) = -\nu (\operatorname{tr} \eta_h, v) + (\det \eta_h - \alpha^2 f, v), \quad \forall v \in V_h \cap H_0^1(\Omega)$$

$$(3.6) \quad T_1(w_h, \eta_h) = w_h \quad \text{on} \quad \partial\Omega.$$

Note that (3.4) is equivalent to

$$(3.7) \quad (T_2(w_h, \eta_h), \tau) + (\operatorname{div} \tau, D T_1(w_h, \eta_h)) - \langle D T_1(w_h, \eta_h), \tau n \rangle = 0 \quad \forall \tau \in \Sigma_h.$$

Let  $I$  denote the  $2 \times 2$  identity matrix. We first make the following important observation.

For  $v \in V_h \cap H_0^1(\Omega)$  and  $\tau = vI$ , we have  $\operatorname{div} \tau = Dv$  and since  $v = 0$  on  $\partial\Omega$ , we have in addition  $\tau n = 0$  on  $\partial\Omega$ . Thus using (3.7) we obtain

$$(3.8) \quad -\nu (\operatorname{tr} T_2(w_h, \eta_h), v) = -\nu (T_2(w_h, \eta_h), vI) = \nu (D T_1(w_h, \eta_h), Dv).$$

Similarly, we obtain that if  $(w_h, \eta_h)$  solves (3.1), then

$$(3.9) \quad (\operatorname{tr} \eta_h, v) = -(D w_h, Dv), \quad \forall v \in V_h \cap H_0^1(\Omega).$$

**Lemma 3.1.** *The mapping  $T$  is well defined and if  $(\alpha w_h, \alpha \eta_h)$  is a fixed point of (3.4)–(3.6) with  $w_h = g_h$  on  $\partial\Omega$ , then  $(w_h, \eta_h)$  solves the nonlinear problem (2.2).*

*Proof.* To prove the first assertion, it is enough to prove that if  $(w_h, \eta_h) \in V_h \times \Sigma_h$  is such that  $w_h = 0$  on  $\partial\Omega$  and

$$\begin{aligned} (\eta_h, \tau) + (\operatorname{div} \tau, Dw_h) - \langle Dw_h, \tau n \rangle &= 0, \forall \tau \in \Sigma_h \\ -\nu(\operatorname{tr} \eta_h, v) &= 0, \forall v \in V_h \cap H_0^1(\Omega), \end{aligned}$$

then  $w_h = 0$  and  $\eta_h = 0$ .

Using (3.9), we obtain  $0 = -(\operatorname{tr} \eta_h, v) = (Dw_h, Dv)$ , for all  $v \in V_h \cap H_0^1(\Omega)$ . Thus  $|w_h|_{H^1}^2 = 0$ . This proves that  $w_h = 0$  by Poincaré's inequality. Using  $\tau = \eta_h$  we obtain as well  $\eta_h = 0$ .

The proof of the second assertion is immediate.  $\square$

We recall from [4, Remark 3.6], see also [9, 8], that for  $v_h \in V_h$ , there exists a unique  $\eta_h \in \Sigma_h$  denoted  $H(v_h)$ , such that

$$(3.10) \quad (H(v_h), \tau) + (\operatorname{div} \tau, Dv_h) - \langle Dv_h, \tau n \rangle = 0, \forall \tau \in \Sigma_h,$$

holds. To see this consider the problem: find  $\eta_h \in \Sigma_h$  such that

$$(3.11) \quad (\eta_h, \tau) = -(\operatorname{div} \tau, Dv_h) + \langle Dv_h, \tau n \rangle, \quad \forall \tau \in \Sigma_h.$$

For  $\tau \in \Sigma_h$ , we define  $F(\tau) = -(\operatorname{div} \tau, Dv_h) + \langle Dv_h, \tau n \rangle$ . Clearly  $F$  is linear. By the Schwarz inequality, (2.5) and (2.7)

$$\begin{aligned} |-(\operatorname{div} \tau, Dv_h) + \langle Dv_h, \tau \cdot n \rangle| &\leq C\|\tau\|_{H^1}\|v_h\|_{H^1} + C\|v_h\|_{H^1(\partial\Omega)}\|\tau\|_{L^2(\partial\Omega)} \\ &\leq C(h^{-1}\|v_h\|_{H^1} + h^{-\frac{1}{2}}\|v_h\|_{H^1(\partial\Omega)})\|\tau\|_{L^2}. \end{aligned}$$

Thus a unique solution  $\eta_h = H(v_h)$  exists by the Lax-Milgram Lemma.

**Remark 3.2.** From the definition of  $H(v_h)$  (3.10) and (3.11), we have for  $v_h \in V_h$ ,

$$H(\alpha v_h) = \alpha H(v_h).$$

**Lemma 3.3.** Let  $v_h \in V_h$  such that  $\|v_h - I_h u\|_{H^1} \leq \mu$ . Then

$$\|H(v_h) - I_h \sigma\|_{L^2} \leq Ch^{-1}\mu + Ch^{k-1}.$$

*Proof.* For  $\tau \in \Sigma_h$ , by (2.1) and (3.10) we have

$$\begin{aligned} (H(v_h) - I_h \sigma, \tau) &= (H(v_h) - \sigma, \tau) + (\sigma - I_h \sigma, \tau) \\ &= (\sigma - I_h \sigma, \tau) - (\operatorname{div} \tau, D(v_h - u)) + \langle D(v_h - u), \tau n \rangle \\ &= (\sigma - I_h \sigma, \tau) - (\operatorname{div} \tau, D(v_h - I_h u)) + \langle D(v_h - I_h u), \tau n \rangle \\ &\quad - (\operatorname{div} \tau, D(I_h u - u)) + \langle D(I_h u - u), \tau n \rangle. \end{aligned}$$

Let  $\tau = H(v_h) - I_h \sigma$ . By the Schwarz inequality, (2.5) and (2.7)

$$\begin{aligned} \|\tau\|_{L^2}^2 &\leq \|\sigma - I_h \sigma\|_{L^2}\|\tau\|_{L^2} + C\|\tau\|_{H^1}\|D(v_h - I_h u)\|_{L^2} \\ &\quad + C\|D(v_h - I_h u)\|_{L^2(\partial\Omega)}\|\tau\|_{L^2(\partial\Omega)} + C\|\tau\|_{H^1}\|D(I_h u - u)\|_{L^2} \\ &\quad + C\|D(I_h u - u)\|_{L^2(\partial\Omega)}\|\tau\|_{L^2(\partial\Omega)} \\ &\leq \|\sigma - I_h \sigma\|_{L^2}\|\tau\|_{L^2} + Ch^{-1}\mu\|\tau\|_{L^2} + Ch^{-1}\|D(v_h - I_h u)\|_{L^2(\Omega)}\|\tau\|_{L^2(\Omega)} \\ &\quad + Ch^{-1}\|\tau\|_{L^2}\|I_h u - u\|_{H^1} + Ch^{-\frac{1}{2}}\|D(I_h u - u)\|_{L^2(\partial\Omega)}\|\tau\|_{L^2}. \end{aligned}$$

Therefore

$$\begin{aligned} \|\tau\|_{L^2} &\leq Ch^{k+1} + Ch^{-1}\mu + Ch^{k-1} + Ch^{k-\frac{1}{2}} \\ &\leq Ch^{-1}\mu + Ch^{k-1}. \end{aligned}$$

This proves the result.  $\square$

It follows from Lemma 3.3, with  $\mu = 0$ , that  $(I_h u, H(I_h u)) \in B_h(\rho)$ , i.e. the ball  $B_h(\rho) \neq \emptyset$  for  $\rho = C_0 h^k$  for a constant  $C_0 > 0$ . See also [4, Lemma 3.5]. As a consequence, see also [9],

$$(3.12) \quad \|H(I_h u) - I_h \sigma\|_{L^2} \leq C_0 h^{k-1}.$$

Let

$$\tilde{B}_h(\rho) = \{v_h \in V_h, v_h = g_h \text{ on } \partial\Omega, \|v_h - I_h u\|_{H^1} \leq \rho\},$$

and consider the mapping

$$\tilde{T}_1 : V_h \rightarrow V_h, \text{ defined by } \tilde{T}_1(v_h) = T_1(v_h, H(v_h)).$$

The motivation to introduce a discrete Hessian  $H(v_h)$  in this paper, as opposed to the approach in [4], is given by Lemma 3.4 below.

**Lemma 3.4.** *If  $w_h$  is a fixed point of  $\tilde{T}_1$ , then  $(w_h, H(w_h))$  is a fixed point of  $T$  and equivalently, if  $(w_h, \eta_h)$  is a fixed point of  $T$ , then  $w_h$  is a fixed point of  $\tilde{T}_1$ .*

*Proof.* The result was given as [4, Remark 3.6]. Let  $w_h$  be a fixed point of  $\tilde{T}_1$ . We have  $T_1(w_h, H(w_h)) = w_h$  and by (3.7) and (3.10),  $T_2(w_h, H(w_h)) = H(T_1(w_h, H(w_h))) = H(w_h)$ . This proves that  $(w_h, H(w_h))$  is a fixed point of  $T$ .

Conversely if  $(w_h, \eta_h)$  is a fixed point of  $T$ , then  $\tilde{T}_1(w_h) = T_1(w_h, H(w_h)) = T_1(w_h, \eta_h) = w_h$ . This completes the proof.  $\square$

**Lemma 3.5.** *We have for  $0 \leq \alpha \leq 1$*

$$(3.13) \quad \|\alpha I_h u - T_1(\alpha I_h u, H(\alpha I_h u))\|_{H^1} \leq \frac{C_1}{\nu} \alpha^2 h^{k-1},$$

for a positive constant  $C_1$ .

*Proof.* Since  $T_1(\alpha I_h u, H(\alpha I_h u)) - \alpha I_h u = 0$  on  $\partial\Omega$ , by (3.8) and (3.5) we have using  $w_h = \alpha I_h u$ ,  $\eta_h = H(\alpha I_h u)$  and  $v = T_1(w_h, \eta_h) - w_h$

$$\nu(DT_1(w_h, \eta_h), Dv) = -\nu(\text{tr } T_2(w_h, \eta_h), v) = -\nu(\text{tr } \eta_h, v) + (\det \eta_h - \alpha^2 f, v).$$

It follows that

$$\nu|Dv|_{L^2}^2 = -\nu(Dw_h, Dv) - \nu(\text{tr } \eta_h, v) + (\det \eta_h - \alpha^2 f, v).$$

Therefore, using (3.9), we get

$$(3.14) \quad \nu|Dv|_{L^2}^2 = (\det \eta_h - \alpha^2 f, v).$$



On the other hand since  $f = \det D^2u = \det \sigma$ , by (2.8) and Remark 3.2, on each element  $K$

$$\begin{aligned}
 \det \eta_h - \alpha^2 f &= \det H(\alpha I_h u) - \alpha^2 \det \sigma = \det \alpha H(I_h u) - \alpha^2 \det \sigma \\
 (3.15) \quad &= \alpha^2 (\det H(I_h u) - \det \sigma) \\
 &= \alpha^2 (\operatorname{cof}(tH(I_h u) + (1-t)\sigma) : (H(I_h u) - \sigma)),
 \end{aligned}$$

for some  $t \in [0, 1]$ .

By (2.3) we have  $\|I_h \sigma\|_{L^\infty} \leq C \|\sigma\|_{L^\infty}$ . Thus by (3.12) and (2.4)

$$\begin{aligned}
 \|H(I_h u)\|_{L^\infty} &\leq \|H(I_h u) - I_h \sigma\|_{L^\infty} + \|I_h \sigma\|_{L^\infty} \leq Ch^{-1} \|H(I_h u) - I_h \sigma\|_{L^2} + \|I_h \sigma\|_{L^\infty} \\
 &\leq Ch^{k-2} + C \|\sigma\|_{L^\infty} \leq C, \text{ since } k \geq 2.
 \end{aligned}$$

Thus by (2.9) and (3.12)

$$\begin{aligned}
 \|\det(H(I_h u)) - \det \sigma\|_{L^2(K)} &\leq C \|tH(I_h u) + (1-t)\sigma\|_{L^\infty(K)} \|H(I_h u) - \sigma\|_{L^2(K)} \\
 &\leq C \|H(I_h u) - \sigma\|_{L^2(K)} \\
 &\leq C \|H(I_h u) - I_h \sigma\|_{L^2(K)} + C \|I_h \sigma - \sigma\|_{L^2(K)} \\
 &\leq Ch^{k-1}.
 \end{aligned}$$

Therefore by (2.3) and (3.15)

$$(3.16) \quad \|\det \eta_h - \alpha^2 f\|_{L^2} = \alpha^2 \|\det(H(I_h u)) - \det \sigma\|_{L^2} \leq C \alpha^2 h^{k-1}.$$

And so combining (3.14)–(3.16), (3.12), Cauchy-Schwarz inequality, the interpolation error estimate (2.3) and Poincaré's inequality, we get

$$|v|_{H^1}^2 \leq \frac{C}{\nu} \alpha^2 h^{k-1} \|v\|_{L^2} \leq \frac{C}{\nu} \alpha^2 h^{k-1} \|v\|_{H^1},$$

from which (3.13) follows. □

We will need the following lemma

**Lemma 3.6.** *Let  $(w_h, \eta_h) \in Z_h$ . Then for a piecewise smooth symmetric matrix field  $P$*

$$(3.17) \quad ((\operatorname{cof} P) : \eta_h, v) + ((\operatorname{cof} P) Dw_h, Dv) \leq Ch \|v\|_{H^1} \|w_h\|_{H^1},$$

for all  $v \in V_h \cap H_0^1(\Omega)$  and for a constant  $C$  which depends on  $\|\operatorname{cof} P\|_{H^{k+1}(\mathcal{T}_h)}$ .

*Proof.* The proof is the same as the proof of [4, Lemma 3.7]. There the proof was given for  $P = D^2u$ , but it carries over to the general case of this lemma line by line. The dependence of the constant  $C$  on  $\|\operatorname{cof} P\|_{H^{k+1}(\mathcal{T}_h)}$  arises from the use in the proof of the approximation property  $\|P_{\Sigma_h}(v \operatorname{cof} P) - v \operatorname{cof} P\|_{H^m(\mathcal{T}_h)} \leq Ch^{k+1-m} \|v \operatorname{cof} P\|_{H^{k+1}(\mathcal{T}_h)}$ . Here  $P_{\Sigma_h}$  denotes the  $L^2$  projection operator into  $\Sigma_h$ . □

**Lemma 3.7.** *For  $(w_h, \eta_h) \in B_h(\rho)$ ,  $\rho = C_0 h^k$ , we have*

$$\|\eta_h - D^2 w_h\|_{L^\infty} \leq Ch^{k-2}.$$

*Proof.* Recall that for  $(w_h, \eta_h) \in B_h(\rho)$ , we have  $\eta_h = H(w_h)$ . We have by (2.4), (3.12)

$$\begin{aligned}
\|\eta_h - D^2 w_h\|_{L^\infty} &\leq \|H(w_h) - D^2 w_h\|_{L^\infty} \\
&\leq \|H(w_h) - I_h \sigma\|_{L^\infty} + \|I_h \sigma - D^2 w_h\|_{L^\infty} \\
&\leq Ch^{-1} \|H(w_h) - I_h \sigma\|_{L^2} + \|I_h \sigma - D^2 u\|_{L^\infty} + \|D^2 u - D^2 w_h\|_{L^\infty} \\
&\leq Ch^{k-2} + Ch^{k+1} + \|D^2 u - D^2 I_h u\|_{L^\infty} + \|D^2 I_h u - D^2 w_h\|_{L^\infty} \\
&\leq Ch^{k-2} + Ch^{-1} \|I_h u - w_h\|_{H^1} \\
&\leq Ch^{k-2}.
\end{aligned}$$

□

The next lemma states a crucial contraction property of the mapping  $T_1$  in  $\alpha B_h(\rho)$ .

**Lemma 3.8.** *Let  $(w_1, \eta_1), (w_2, \eta_2) \in B_h(\rho)$  with  $\rho \leq \min(C_0, C_{conv})h^k$ . We have*

$$(3.18) \quad |T_1(\alpha w_1, \alpha \eta_1) - T_1(\alpha w_2, \alpha \eta_2)|_{H^1} \leq a |\alpha w_1 - \alpha w_2|_{H^1},$$

for  $0 < a < 1$ ,  $h$  sufficiently small,  $\alpha = h^{k+2}$  and  $\nu = (m + M)/2$ .

*Proof.* Put  $v = T_1(\alpha w_1, \alpha \eta_1) - T_1(\alpha w_2, \alpha \eta_2)$ . By assumption  $v \in V_h \cap H_0^1(\Omega)$ . Using (3.8) and (3.5) we obtain

$$\begin{aligned}
\nu(DT_1(\alpha w_1, \alpha \eta_1) - DT_1(\alpha w_2, \alpha \eta_2), Dv) &= -\nu(\text{tr } T_2(\alpha w_1, \alpha \eta_1) - \text{tr } T_2(\alpha w_2, \alpha \eta_2), v) \\
&= -\nu(\text{tr } \alpha \eta_1 - \text{tr } \alpha \eta_2, v) + (\det \alpha \eta_1 - \det \alpha \eta_2, v).
\end{aligned}$$

Therefore, using (2.8), we have for some  $t \in [0, 1]$  and with the notation

$$Q = t\eta_1 + (1-t)\eta_2 \text{ and } \overline{Q} = tD^2 w_1 + (1-t)D^2 w_2,$$

$$\begin{aligned}
|v|_{H^1}^2 &= -(\text{tr } \alpha \eta_1 - \text{tr } \alpha \eta_2, v) \\
&\quad + \frac{1}{\nu}((\text{cof } \alpha(t\eta_1 + (1-t)\eta_2)) : \alpha(\eta_1 - \eta_2), v) \\
&= \left((-I + \frac{1}{\nu} \text{cof } \alpha Q) : \alpha(\eta_1 - \eta_2), v\right) \\
(3.19) \quad &= -(I : \alpha(\eta_1 - \eta_2), v) - (D\alpha(w_1 - w_2), Dv) \\
&\quad + \frac{1}{\nu}((\text{cof } \alpha Q) : \alpha(\eta_1 - \eta_2), v) + \frac{1}{\nu}((\text{cof } \alpha Q) D\alpha(w_1 - w_2), Dv) \\
&\quad + (D\alpha(w_1 - w_2), Dv) - \frac{1}{\nu}((\text{cof } \alpha \overline{Q}) D\alpha(w_1 - w_2), Dv) \\
&\quad + \frac{1}{\nu}((\text{cof } \alpha \overline{Q}) D\alpha(w_1 - w_2), Dv) - \frac{1}{\nu}((\text{cof } \alpha Q) D\alpha(w_1 - w_2), Dv).
\end{aligned}$$

For  $(w_1, \eta_1), (w_2, \eta_2) \in B_h(\rho)$ ,  $t(w_1, \eta_1) + (1-t)(w_2, \eta_2) \in B_h(\rho)$  and thus for  $h$  sufficiently small, by Lemmas 2.2 and 2.3 we get

$$(3.20) \quad |(D(w_1 - w_2), Dv) - \frac{1}{\nu}((\text{cof } \alpha \overline{Q}) D(w_1 - w_2), Dv)| \leq \gamma |w_1 - w_2|_{H^1} |v|_{H^1},$$

for  $0 < \gamma < 1$ .

On the other hand, by Lemma 3.6, with  $P = I$ , we have

$$(3.21) \quad |-(I : (\eta_1 - \eta_2), v) - (D(w_1 - w_2), Dv)| \leq Ch|w_1 - w_2|_{H^1}|v|_{H^1}.$$

Applying Lemma 3.6, with  $P = Q$ , we get

$$(3.22) \quad |((\operatorname{cof} Q) : (\eta_1 - \eta_2), v) + ((\operatorname{cof} Q)D(w_1 - w_2), Dv)| \leq Ch\|\operatorname{cof} Q\|_{H^{k+1}(\mathcal{T}_h)}|w_1 - w_2|_{H^1}|v|_{H^1}.$$

Finally, since by (2.10)

$$\operatorname{cof} Q - \operatorname{cof} \bar{Q} = \operatorname{cof}(Q - \bar{Q}) = \operatorname{cof}\left(t(\eta_1 - D^2 w_1) + (1 - t)(\eta_2 - D^2 w_2)\right),$$

we get using Lemma 3.7

$$\|\operatorname{cof} Q - \operatorname{cof} \bar{Q}\|_{L^\infty} \leq Ch^{k-2} \leq C, \text{ since } k \geq 2.$$

Thus

$$(3.23) \quad \left| \frac{1}{\nu}((\operatorname{cof} \bar{Q})D(w_1 - w_2), Dv) - \frac{1}{\nu}((\operatorname{cof} Q)D(w_1 - w_2), Dv) \right| \leq C|w_1 - w_2|_{H^1}|v|_{H^1}.$$

We conclude from (3.19)–(3.23) that

$$(3.24) \quad |v|_{H^1} \leq (\gamma + Ch + C\alpha h\|\operatorname{cof} Q\|_{H^{k+1}(\mathcal{T}_h)} + C\alpha)|\alpha w_1 - \alpha w_2|_{H^1}.$$

Using the inverse estimate (2.6) and noting that  $\rho \leq h^2$

$$\begin{aligned} \|\operatorname{cof} Q\|_{H^{k+1}(\mathcal{T}_h)} &\leq Ch^{-k-1}\|\operatorname{cof} Q\|_{L^2} \leq Ch^{-k-1}\|Q\|_{L^2} \\ &\leq Ch^{-k-1}\|t\eta_1 + (1 - t)\eta_2\|_{L^2} \\ &\leq Ch^{-k-1}(\|\eta_1\|_{L^2} + \|\eta_2\|_{L^2}) \\ &\leq Ch^{-k-1}(\|\eta_1 - I_h\sigma\|_{L^2} + \|\eta_2 - I_h\sigma\|_{L^2} + 2\|I_h\sigma\|_{L^2}) \\ &\leq Ch^{-k-1}(h^{-1}\rho + \|\sigma\|_{L^2}) \leq Ch^{-k-1}(Ch + \|\sigma\|_{L^2}) \leq Ch^{-k-1}. \end{aligned}$$

Since  $\gamma < 1$ , and  $\alpha = h^{k+2}$ , for  $h$  sufficiently small,  $Ch + C\alpha h\|\operatorname{cof} Q\|_{H^{k+1}(\mathcal{T}_h)} + C\alpha < 1 - \gamma$ . We conclude from (3.24) that (3.18) holds.  $\square$

**Lemma 3.9.** *For  $\rho = \min(C_0, C_{conv})h^k$ , the mapping  $\tilde{T}_1$  has a unique fixed point in  $\alpha\tilde{B}_h(\rho)$  for  $\alpha = h^{k+2}$ .*

*Proof.* Note that by (3.18),  $\tilde{T}_1$  is a strict contraction in  $\alpha\tilde{B}_h(\rho)$  for  $\rho \leq \min(C_0, C_{conv})h^k$ . We now show that  $\tilde{T}_1$  maps  $\alpha\tilde{B}_h(\rho)$  into itself. Let  $v_h \in \tilde{B}_h(\rho)$ . We have by (3.18) and (3.13)

$$\begin{aligned} \|\tilde{T}_1(\alpha v_h) - \alpha I_h u\|_{H^1} &\leq \|\tilde{T}_1(\alpha v_h) - \tilde{T}_1(\alpha I_h u)\|_{H^1} + \|\tilde{T}_1(\alpha I_h u) - \alpha I_h u\|_{H^1} \\ &\leq a\|\alpha v_h - \alpha I_h u\|_{H^1} + C_1\alpha^2 h^{k-1} \\ &\leq a\alpha\rho + C_1\alpha h^{2k+1} = a\alpha\rho + C_1 h^{k+1}\alpha h^k. \end{aligned}$$

Therefore for  $h$  sufficiently small,  $C_1 h^{k+1} \leq \min(C_0, C_{conv})(1 - a)$  and so

$$\|\tilde{T}_1(\alpha v_h) - \alpha I_h u\|_{H^1} \leq a\alpha\rho + (1 - a)\alpha\rho.$$

The result then follows from the Banach fixed point theorem.  $\square$

We can now state the main result of this paper

**Theorem 3.10.** *Problem (2.2) has a unique local solution  $(u_h, \sigma_h)$  for  $k \geq 2$  and  $h$  sufficiently small. We have*

$$\begin{aligned} \|u_h - I_h u\|_{H^1} &\leq Ch^k \\ \|\sigma_h - I_h \sigma\|_{H^1} &\leq Ch^{k-1}. \end{aligned}$$

*Proof.* Recall that for  $(u_h, \sigma_h) \in B_h(\rho)$ , we have  $\sigma_h = H(u_h)$ . The result follows from Lemmas 3.4, 3.9 and 3.1, the definition of  $B_h(\rho)$  and (3.12).

The local solution  $u_h$  given by Lemma 3.9 satisfies  $\|u_h - I_h u\|_{H^1} \leq Ch^k$ . Since by Lemma 3.4,  $(u_h, H(u_h))$  is a fixed point of  $T$ , by Lemma 3.1,  $(u_h, H(u_h))$  solves (2.2). By the definition of  $B_h(\rho)$   $\sigma_h = H(u_h)$  and by (3.12), we have  $\|\sigma_h - I_h \sigma\|_{H^1} \leq Ch^{k-1}$ . □

## REFERENCES

- [1] Awanou, G.: Pseudo transient continuation and time marching methods for Monge-Ampère type equations (2013). <http://arxiv.org/pdf/1301.5891.pdf>
- [2] Awanou, G.: On standard finite difference discretizations of the elliptic Monge-Ampère equation (2014). Submitted
- [3] Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equation: classical solutions (2014). To appear in IMA J. of Num. Analysis
- [4] Awanou, G., Li, H.: Error analysis of a mixed finite element method for the Monge-Ampère equation. Int. J. Num. Analysis and Modeling **11**, 745–761 (2014)
- [5] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. I. Math. Comp. **47**(175), 103–134 (1986)
- [6] Brenner, S.C., Gudi, T., Neilan, M., Sung, L.Y.:  $C^0$  penalty methods for the fully nonlinear Monge-Ampère equation. Math. Comp. **80**(276), 1979–1995 (2011)
- [7] Brenner, S.C., Scott, L.R.: The mathematical theory of finite element methods, *Texts in Applied Mathematics*, vol. 15, second edn. Springer-Verlag, New York (2002)
- [8] Lakkis, O., Pryer, T.: A finite element method for nonlinear elliptic problems. SIAM J. Sci. Comput. **35**(4), A2025–A2045 (2013)
- [9] Neilan, M.: Finite element methods for fully nonlinear second order PDEs based on a discrete Hessian with applications to the Monge-Ampère equation. J. Comput. Appl. Math. **263**, 351–369 (2014)

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE, M/C 249. UNIVERSITY OF ILLINOIS AT CHICAGO, CHICAGO, IL 60607-7045, USA

*E-mail address:* awanou@uic.edu

*URL:* <http://www.math.uic.edu/~awanou>